STATE-SPACE-LIKE MODELS TO CALL COPY NUMBERS

Ellen Visccher & Christopher Yau
Nuffield Department of Women's and Reproductive Health, University of Oxford
{ellen.visscher, christopher.yau}@{bdi, wrh}.ox.ac.uk

1 INTRODUCTION

Somatic copy number alterations (CNAs) are genomic regions amplified or deleted during somatic cell replication, playing a critical role in cancer development by driving oncogene amplification (Zhang & Pellman, 2022; Kim et al., 2020; Rosswog et al., 2021). Accurate CNA profiling is essential for downstream analysis. Despite advances in 'omics and deep learning, CNA callers have seen limited innovation, due to sequence length limitations of transformers. To address this, we propose araCNA, a deep learning-based approach to improve CNA calling, leveraging state-space-like models that enable genome-scale modeling.

2 METHODS

We first define the mathematical construction, similar to that first shown in (Van Loo et al., 2010). We define $C_{T,i}$, $C_{P,i}$, $C_{M,i}$, $C_{B,i}$ as the total, paternal, maternal and B-allele copy number at locus i in the tumour. We further define ρ as the purity of the tumour sample (the proportion of tumour vs non-tumour) and r_d as the expected number of sequencing reads per copy number, which are both unknown. Finally, we have R_i, the total number of reads at a locus and B_i, the B allele frequency (BAF), both measured data. The sequence data is therefore a collection $\{R_i, B_i\}_{i=1}^{L}$ for L loci. The generating process from $\{C_{M,i}, C_{P,i}, r_d, \rho\} \to \{R_i, B_i\}$ is well understood, Figure S1, Appendix A.1, what remains difficult is the inference process, $\{R_i, B_i\} \to \{A_{M,i}, A_{m,i}, r_d, \rho\}$. Here major and minor parental copy numbers, $A_M = \max(C_M, C_P)$ and $A_m = \min(C_M, C_P)$, are introduced due to the non-identifiability of (C_M, C_P).

The function of araCNA can be summarised as $f_\theta(\{R_i, B_i\}) \to (\{p_{i,k}\}, \hat{\rho})$ where f_θ is a state-space- like model (i.e Hyena, (Poli et al., 2023) or Mamba, (Gu & Dao, 2024)) parameterised by network weights θ, Figure 1. $\hat{\rho}$ is araCNA's global purity estimate. While $p_{k,i}$ is the probability that araCNA assigns the locus as belonging to copy number profile K_k. The profile categories $K_1, \ldots, K_J$ correspond to major/minor parental copy number combinations. From these, we can also estimate $\hat{r}_d$, Appendix A.1.

We trained our model using simulated datasets where the ground truth copy numbers and purity are known. The loss function consists of a supervised sequence loss for copy numbers and a global loss for purity and read depth predictions. The model was trained by iteratively increasing the complexity of the simulated data. For details see Appendix A.2. Figure 2 demonstrates the complexity of real data samples that araCNA learns to infer.

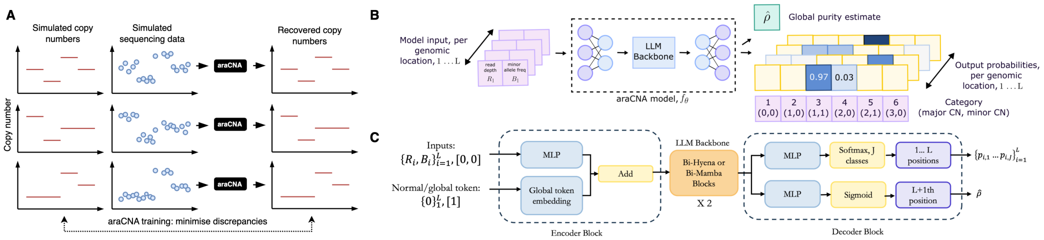

Figure 1: Overview of A) training of araCNA B) high-level model, C) araCNA architecture.

3 RESULTS

We first compared results from our two araCNA variants (araCNA-mamba and araCNA-hyena) using simulated data, Figure 3A, simulation procedure in Appendix A.3.

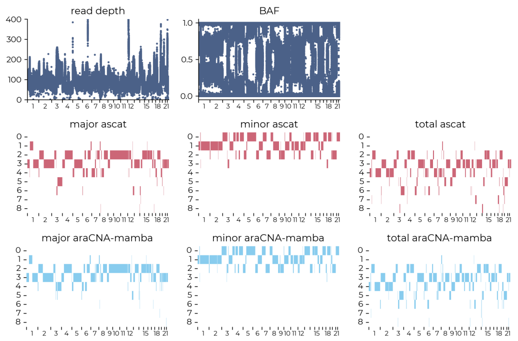

Figure 2: Representative TCGA ovarian cancer sample. Comparison of predicted profiles from ASCAT and araCNA-mamba

Both models achieve high copy number classification accuracy for the task though araCNA-mamba slightly outperforms araCNA-hyena. In this case, we know the ground truth so can directly measure accuracy.

We next compare results araCNA to several existing popular CNA calling tools for whole genome sequencing data by analysing a selection of 50 tumour samples chosen from the Cancer Genome Atlas (TCGA). Since there are no ground truth copy number profiles for these tumours, we compare methods to each other, using concordance Figure 3B, and use proxy measures such as reconstruction error to provide an unsupervised metric for performance, where better methods will be expected to have low reconstruction error Figure 3C. However, a low reconstruction error accompanied by a large number of segments may suggest overfitting, while high copy number calls at high read depth regions will have a lower reconstruction error but are likely less plausible.

Using both araCNA variants for zero-shot inference of the copy number states gives comparable reconstruction performance to the existing CNA calling methods while using similar numbers of segments. ASCAT and Battenberg achieve slightly improved reconstruction error performance, however, assign as high as 100 copies to localised genomic regions,Figure 3C. Conversely, araCNA achieves this performance despite being limited to calling a total copy number of 8.

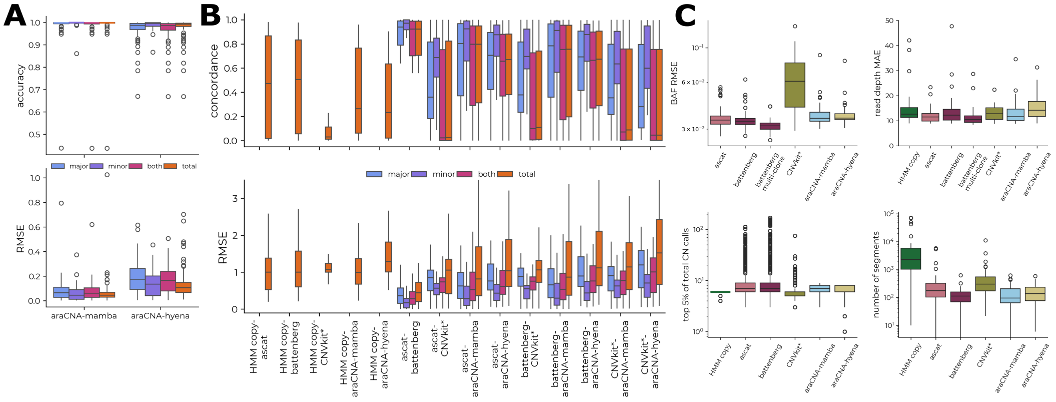

Figure 3: A) Accuracy and RMSE from 100 simulated genomic samples of length 650k, drawn from same distribution as training data. B) Concordance and RMSE between different callers on 50 TCGA cancer samples on WGS data C) Boxplots showing the B-allele-frequency reconstruction RMSE, read-depth mean absolute error (MAE), copy number distribution of top 5% of copy number calls and the number of different copy number segments identified across the 50 samples.

4 Conclusion

Here, we present a novel deep-learning approach, araCNA, trained only on simulated data that can accurately predict CNAs in real WGS cancer genomes. araCNA uses novel transformer alternatives (e.g Mamba) to handle genomic-scale sequence lengths ($\sim$ 1M) and learn long-range interactions. Results are extremely accurate on simulated data, and this zero-shot approach is on par with existing methods when applied to 50 WGS samples from the cancer genome atlas. Notably, our approach requires only a tumour sample and not a matched normal sample, has fewer markers of overfitting, and performs inference in only a few minutes.

Meaningfulness Statement

What is more literal to "life" than DNA. Here we seek to uncover the true representation of regions of patient DNA that have been amplified or deleted due to cancerous aberrations, measured using whole genome sequencing (WGS). Unlike traditional methods, we use novel deep-learning architectures that learn to represent (an analogue of) the posterior distribution of copy number profiles from WGS data. We use simulated data to circumvent the unknown ground-truth problem that affects much of biology. Our approach demonstrates a proof-of-concept for biological applications where the generating process is understood but inference is challenging.

Acknowledgments

CY is supported by an UKRI Turing AI Acceleration Fellowship (Ref: EP/V023233/1) and EPSRC grant (Ref: EP/Y018192/1). EV is supported by the Oxford EPSRC Centre for Doctoral Training in Health Data Science (Ref: EP/S02428X/1).

References

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, May 2024. URL http://arxiv.org/abs/2312.00752. arXiv:2312.00752 [cs].

Hoon Kim, Nam-Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D. Gujar, Jens Luebeck, Jihe Liu, Viraj Deshpande, Utkrisht Rajkumar, Sandeep Namburi, Samirkumar B. Amin, Eunhee Yi, Francesca Menghi, Johannes H. Schulte, Anton G. Henssen, Howard Y. Chang, Christine R. Beck, Paul S. Mischel, Vineet Bafna, and Roel G. W. Verhaak. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nature Genetics*, 52(9): 891–897, September 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-0678-2. URL https://www.nature.com/articles/s41588-020-0678-2. Publisher: Nature Publishing Group.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *Advances in Neural Information Processing Systems*, 36:43177–43201, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Re. Hyena Hierarchy: Towards Larger Convolutional Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28043–28078. PMLR, July 2023. URL https://proceedings.mlr.press/v202/poli23a.html. ISSN: 2640-3498.

Carolina Rosswog, Christoph Bartenhagen, Anne Welte, Yvonne Kahlert, Nadine Hemstedt, Witali Lorenz, Maria Cartolano, Sandra Ackermann, Sven Perner, Wenzel Vogel, Janine Altmüller, Peter Nürnberg, Falk Hertwig, Gudrun Göhring, Esther Lilienweiss, Adrian M. Stütz, Jan O. Korbel, Roman K. Thomas, Martin Peifer, and Matthias Fischer. Chromothripsis followed by circular recombination drives oncogene amplification in human cancer. *Nature Genetics*, 53(12):1673–1685, December 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00951-7. URL https://www.nature.com/articles/s41588-021-00951-7. Publisher: Nature Publishing Group.

Peter Van Loo, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M. Perou, Anne-Lise Børresen-Dale, and Vessela N. Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, September 2010. Publisher: Proceedings of the National Academy of Sciences.

Cheng-Zhong Zhang and David Pellman. Cancer Genomic Rearrangements and Copy Number Alterations from Errors in Cell Division. *Annual Review of Cancer Biology*, 6(Volume 6, 2022):245–268, April 2022. ISSN 2472-3428. doi: 10.1146/annurev-cancerbio-070620-094029. URL `https://www.annualreviews.org/content/journals/10.1146/annurev-cancerbio-070620-094029`. Publisher: Annual Reviews.

A Appendix

A.1 Generating process

For a pure tumour sample, we have the total copy number as:

$$C_{T,i} = C_{P,i} + C_{M,i}$$

Considering sample impurity, we define the sample copy number C^s_T as:

$$C^s_{T,i} = \rho C_{T,i} + 2(1-\rho),$$

where we assume the contaminating normal cells have copy number 2 at all loci. While normal cells may possess some copy number variants, the size of these regions are typically negligible compared to the cancer-associated alterations we aim to detect and so we ignore these for simplicity.

The B allele copy number is defined as:

$$C_{B,i} = s_{p,i}C_{P,i} + s_{m,i}C_{M,i}$$

where $(s_{p,i}, s_{m,i}) \in \{(0,0),(0,1),(1,0),(1,1)\}$ denotes whether the paternal and maternal chromosomes respectively have the specified SNP B allele at locus i.

Adding sample impurity we have:

$$C^s_{B,i} = s_{p,i}((1-\rho) + \rho C_{P,i}) + s_{m,i}((1-\rho) + \rho C_{M,i})$$

The total number of reads at a locus R_i are then:

$$R_i = r_d C^s_{T,i}$$

While the B allele frequency (BAF) B_i at locus i is given by:

$$B_i = \frac{C^s_{B,i}}{C^s_{T,i}}$$

To determine the copy number profiles, we must infer $\hat{}, \{\hat{C}_{M,i}, \hat{C}_{m,i}\}$. From these we can also estimate $\hat{r}_d = \mu_{\text{robust}}(R)/(\mathbb{E}[\hat{C}^s])$, where $\mu_{\text{robust}}(R)$ is the robust or trimmed mean of the read depth vector and $\mathbb{E}[\hat{C}^s]$ is the expected value of the overall sample copy number (ploidy).

Figure S1 illustrates the link between the mathematical construction and measured data from a tissue sample.

A.2 Loss and training procedure

The loss used to train `araCNA` is given by:

$$\mathcal{L}_{ss} = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{J} I\{c_i = K_j\}\log(p_{k,i}) + \lambda_r|r_d - \hat{r}_d| + \lambda_\rho|\rho - \hat{\rho}|$$

where L is the sequence length, $c_i \in K_1 \dots K_J$ is the known target profile of a genomic locus.

The first term is supervised sequence loss is the cross entropy, while the last two terms are supervised global parameter losses. We found $\lambda_r = \lambda_\rho = 1$ to work well.

To train `araCNA` we adopted an iterative warmup procedure, gradually increasing the complexity of the problem. We found this was necessary for the model to learn, and a similar approach was taken in (Nguyen et al., 2023) with gradually increasing the sequence length.

The training procedure was:

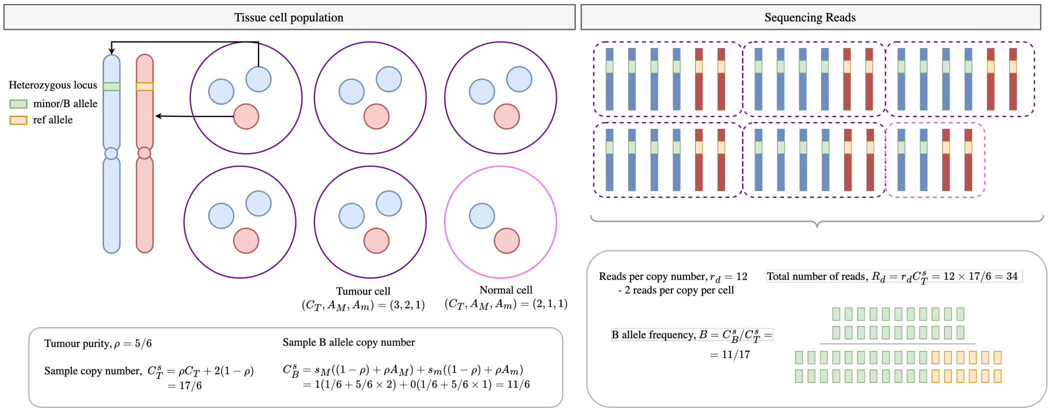
Tissue cell population
Sequencing Reads
Heterozygous locus
minor/B allele
ref allele
Tumour cell
$(C_T, A_M, A_m) = (3, 2, 1)$
Normal cell
$(C_T, A_M, A_m) = (2, 1, 1)$
Tumour purity, $\rho = 5/6$
Sample copy number, $C_T^s = \rho C_T + 2(1-\rho)$
$= 17/6$
Sample B allele copy number
$C_B^s = s_M((1-\rho) + \rho A_M) + s_m((1-\rho) + \rho A_m)$
$= 1(1/6 + 5/6 \times 2) + 0(1/6 + 5/6 \times 1) = 11/6$
Reads per copy number, $r_d = 12$
- 2 reads per copy per cell
Total number of reads, $R_d = r_d C_T^s = 12 \times 17/6 = 34$
B allele frequency, $B = C_B^s / C_T^s =$
$= 11/17$

Figure S1: Mathematical construction of copy number calling. Illustration of how purity, copy number, read depth per-copy number and heterozygous loci (major/minor haplotypes denoted s_M, s_m) result in measured read depth and B allele frequency.

1. Begin the synthetic data generating procedure with $\rho = 1$, and without sampling the noise parameters. Use only up to a maximum total copy number of 2, that is profiles, $(A_M, A_m) \in \{(0,0), (1,0), (1,1), (2,0)\}$. Sample r_d, and start with sequence length 10000. Train until convergence.
2. Using the previously trained model weights as initialisation, add in purity and noise parameter sampling. Train until convergence.
3. Using the previously trained model weights as initialisation, slowly increase the maximum total copy number to 8. Train until convergence.
4. Using the previously trained model weights as initialisation, slowly increase the maximum sequence length to 650,000. Train until convergence.

A.3 Synthetic data simulation

We generated synthetic copy number profiles using the following procedures:

1) Sampling the number of segments. We sample the approximate number of copy number segments, $\hat{N}_s$ using a mixture approach; first, we sample a uniform variable, u such that under a user-defined swap probability, q_s, the number of segments is sampled uniformly between 1 and N. When $u > q_s$, a Poisson distribution is used to skew sampling towards smaller total segments. This is to oversample harder cases with fewer segments where it is harder to estimate global parameters like read depth per copy number and purity.

2) Sampling the segment breakpoints. This is done by randomly sampling $b_1, \ldots, b_{\hat{N}_s}$ breakpoints from $1 \ldots L$, the unique set of these breakpoints defines the segments, and $N_s = |b_1, \ldots, b_{\hat{N}_s}|$. We only keep the segments that have a minimum segment length of L_{min}.

3) Sampling the segment profiles. We sample A_M, A_m of each segment from the possible copy number profiles. We inject logic here to preferentially sample profiles closer in copy number to 1-1 when there are fewer segments. This is due to the identifiability issue. When there are more segments, profiles are sampled more uniformly but still with a preference for lower copy numbers, to inject an implicit bias towards lower ploidy solutions when the model is unsure.

From a sampled profile, we simulate the sequencing read depth and B allele frequency data. Each of the L loci is considered a commonly varying single nucleotide polymorphism, SNP. For both parental alleles, A_M, A_m, we sample each SNP as binomial with a probability of 0.5, we also sample the purity, ρ, uniformly from a range between 0.5 and 1. This gives the sample minor allele copy number, $C^s_{B,i}$ and the sample total copy number, $C^s_{T,i}$. The read depth per copy number, r_d is sampled uniformly between 5 and 70, and together with $C^s_{T,i}$ the overall read depth, R_i is sampled from this mean with additional noise. The BAF, B_i is sampled using total reads sampled based on R_i and the subset of B-allele reads using a binomial probability of $C^s_{B,i}/C^s_{T,i}$, with added noise.

In real data, there exist regions of prolonged homozygosity that can be attributed to identity-by-descent (IBD) regions (i.e identical regions inherited from both parents due to a common ancestor). To emulate this, we also randomly inject regions of prolonged homozygosity into the model. In these IBD regions, the BAF cannot be used to infer copy-number, and the model must use context from before/after the homozygous region for correct prediction.

From this sampling procedure, we therefore have a set of targets $(\{A_{M,i}, A_{m,i}\}, \rho, r_d)$ that generate inputs $\{R_i, B_i\}$, which together are used in the training of araCNA. Hence, araCNA can be interpreted as performing inference on the above statistical approach, when $(\{A_{M,i}, A_{m,i}\}, \rho, r_d)$ are treated as unknowns.